Individual & collective human intelligence in drug design: evaluating the search strategy

Giovanni Cincilla*, Simone Masoni* and Jascha Blobel*

Molomics, Barcelona Science Park, c/ Baldiri i Reixac 4-12, 08028 Barcelona, Spain.

* Corresponding author

E-mail: Giovanni Cincilla - gcincilla@molomics.com; Simone Masoni – smasoni@molomics.com; Jascha Blobel – jblobel@molomics.com

1 Abstract

In recent years, individual and collective human intelligence, defined as the knowledge, skills, reasoning and intuition of individuals and groups, have been used in combination with computer algorithms to solve complex scientific problems. Such approach was successfully used in different research fields such as: structural biology, comparative genomics, macromolecular crystallography and RNA design. Herein we describe an attempt to use a similar approach in small-molecule drug discovery, specifically to drive search strategies of *de novo* drug design. This is assessed with a case study that consists of a series of public experiments in which participants had to explore the huge chemical space *in silico* to find predefined compounds by designing molecules and analyzing the score associate with them. Such a process may be seen as an instantaneous surrogate of the classical design-make-test cycles carried out by medicinal chemists during the drug discovery hit to lead phase but not hindered by long synthesis and testing times. The objectives of this case study are to give the first insights towards: the assessment of human intelligence in chemical space exploration problems; compare the performance of individual and collective human intelligence in such a problems; and also contrast some human and artificial intelligence achievements in *de novo* drug design.

Keywords: collective intelligence, chemical space exploration, *de novo* drug design, artificial intelligence.

2 Introduction

In the last decade, different citizen science initiatives have been promoted to solve complex scientific problems using crowdsourcing and gamification.[1–3] To achieve its objectives, these initiatives make use of individual and collective human intelligence, defined as the knowledge, skills, reasoning and intuition of individuals and groups. Probably the most known projects of this type, developed as on-line video games, are: FoldIt, Phylo, CrowdPhase, Udock and EteRNA. FoldIt predicts protein structures[4–7]and deals with *de novo* protein design;[8] Phylo[9] answers multiple sequence alignment questions of comparative genomics; CrowdPhase[10,11] addresses *ab initio* phasing issues of macromolecular crystallography; Udock[12,13] tackles protein-protein docking puzzles and EteRNA[14,15] solves in vitro

RNA design problems. The commonality of these approaches is that they address complex problems with many degrees of freedom where computational approaches struggle to find optimal solutions between the huge number of possible ones.

In the field of small-molecule drug discovery a problem of this type is represented by the drug design process. Actually, designing an ideal drug corresponds to finding an optimal molecule in the chemical space. This is an extremely hard task *inter alia* because the chemical space is huge and finding a specific molecule therein is a needle-in-a-haystack problem.

The chemical space, defined as that abstract entity containing the sum of all drug-like small-molecules, is awfully large. A rigorous method to estimate its extent doesn't exist. The probably most cited size is 10^{60} different molecules, whereas the real number should be somewhere between 10^{23} and 10^{180}.[16–22] What extent of the chemical space has already been explored? To date: 10^{8} molecules have been already synthesized;[a,b] 10^{11} molecules constitute the largest systematic enumeration of all the synthetically accessible molecules up to 17 atoms;[23] and 10^{13} synthetically accessible molecules can be virtually screened.[24]Although reaching such amounts constitutes certainly a great achievement, this is almost insignificant in respect to the total number of possible molecules.

An efficacious way to explore and exploit the chemical space without the need of enumerating huge amounts of molecules is using *de novo* molecular designers. These are automatic *in silico* techniques that create molecules from scratch, optimizing certain previously defined requirements (i.e. molecular properties).[25] Any *de novo* designer is composed of three elements: a scoring strategy, the method with which molecules are evaluated; an assembly strategy, the approach with which molecules are built; and a search strategy, the technique with which molecules are searched in the chemical space.[26] Many automatic *de novo* systems have been designed, implemented and tested since almost three decades. They use different scoring strategies (i.e. structure-based,[27–29] ligand-based;[30,31] both coupled with single- and multi-objective optimization approaches[32,33]), assembly strategies (e.g. atom/bond-based, fragment-based, reaction-based) and search strategies (e.g. Machine Learning,[34–39]Genetic Algorithms[30,40–42]). Although several of these methods have shown promising results, their validation has not been consistent. To solve this problem a suite of benchmarks for *de novo* molecular design has been recently proposed.[43]

The three constitutional elements of *de novo* designers (i.e. search, assembly and scoring strategies) are not specific of the *in silico* approach but are general characteristics of the molecule design process. Actually, the same components are part of the classical design-make-test optimization cycles used by medicinal chemists in drug discovery with which initial hit molecules are optimized to leads. Indeed *de novo* designers carry out virtual design-make-test cycles *in silico*.

Until today only timid attempts have been made to address drug design using crowdsourcing. Recently

[a] As 27/11/2020 PubChem contains 110,507,961compounds. https://www.ncbi.nlm.nih.gov/pccompound?term=all%5Bfilt%5D&cmd=search

[b] As 27/11/2020 CAS registry contains more than 171 million unique organic and inorganic chemical substances (https://www.cas.org/support/documentation/chemical-substances)

some trials were done by integrating many experts in order to: enhance chemical libraries through the "wisdom of crowds",[44] model molecular complexity from a crowdsourced medicinal chemist perspective[45], predict solubility in place of machines,[46] and assess quality of molecules generated by automatic algorithms in Turing-inspired tests.[47] All such activities are related to scoring strategies of *de novo* drug design but no endeavor has been made (as far as we know) to deal with the other two elements: the assembly and the search strategies.

Herein we describe an attempt to use individual and collective human intelligence as search strategies of *de novo* drug design and quantify their performance. To our knowledge this is the first time that artificial intelligence (e.g. machine learning, genetic algorithms) is substituted by human intelligence in an *in silico, de novo* drug design process.

The case study consisted of a series of public experiments addressed to the scientific community where each participant was asked to explore the chemical space both individually and collectively. The search started from scratch, meaning a single carbon atom, which could be extended and modified to nearly any molecular structure. The final goal was to design molecules that maximize a single-value scoring function. Since this is also the goal of automatic *de novo* drug designers, results obtained by humans were then compared with those obtained by machines. As the first study or its kind, we used a molecular similarity function as a score for the chemical space exploration. This is a typical first step before using more complex, multi-objective functions (e.g. constituted by different machine learning models) that are more suitable for drug discovery programs. In fact molecular similarity is a surrogate for machine learning models and has two big advantages: on one side it is easily interpretable; and on the other side the successful design of the predefined target molecule, towards which the similarity functions achieve their maximum, can be unequivocally determined.

The final objectives of this case study were:

1. Assess human intelligence in chemical space exploration problems
2. Compare individual vs collective human intelligence performances in molecule design
3. Contrast human intelligence with artificial intelligence results obtained in *de novo* drug design

3 Methods

This section describes the methodology used to design, carry out and analyze this case study. Details about the software application used for the study and its implementation are reported in *Supporting Information.*

3.1 Experiment settings & circumstances

The case study[c] consisted of a series of public experiments where each participant should find a specific, predefined target molecule in the chemical space. This was supposed to be done by designing

[c] http://molomics.com/explore

molecules from scratch, following a molecular score that indicates how close the solution is. Participants were invited to engage in two experiments: an individual design and a collective design experiment. In the first one they searched the target molecule individually by competing with other participants, while in the second one they did it collectively by collaborating amongst each other.

The scientific community was invited to take part in this case study through social networks (i.e. Twitter, LinkedIn). An *ad hoc* application was developed to carry out this case study which design, characteristics and implementation are described in *Supporting Information.* Before being invited to the experiments, participants were asked to create an account on our application and undertake simple learning steps in the *Sandbox*, the application area where one can learn how to draw, save and access molecules. Participants that fulfilled the *Sandbox* requirements were consecutively invited to an individual and a collective design experiment. The beginning of an experiment was scheduled only once at least 10 participants were available. At least 24 hours before the experiment started, the participants were notified by an e-mail system which is described in *Supporting Information.* Different experiments could be launched and run at the same time by randomly selecting participants between those who fulfilled the Sandbox requirements. The duration of each experiment was set to the first occurring event, being either the discovery of the target molecule or a time limit of two weeks. None of the participants was involved simultaneously in the two experiments associated to them.

Collective but also individual design experiments were run with groups of people for two main reasons. First, the settings of the two experiment types were supposed to be maintained as similar as possible. Second, in this way participants had access to the experiment common ranking that worked as a motivation factor to drive the molecular search.

From a practical point of view, the main difference between an individual and a collective design experiment is that while in the former a participant has only access to the molecules generated by her/himself, in the latter she/he has access at any moment to all the molecules generated by all the participants of the experiment, dynamically.

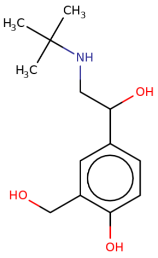
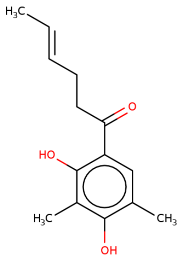
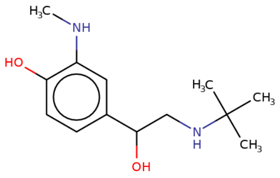
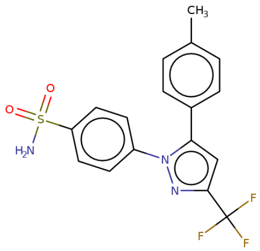
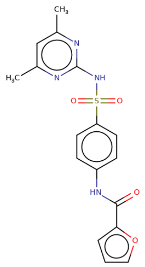
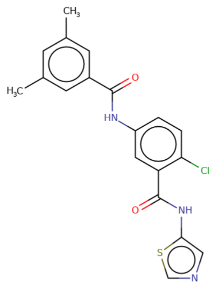
3.2 The target molecules

In order to assess the human capacity of exploring the chemical space but also compare it to that of automatic *de novo* methods, five benchmarks were selected from a recently published benchmark suite[43] for *de novo* drug design. As explained in section 3.5, these benchmarks are based on five target molecules of five different complexity levels. For each of these complexity levels, one individual and one collective design experiment were planned, resulting in a total of 10 experiments.

Nevertheless, using the published target molecules of the five selected benchmarks with humans may bring to potential disputes. First, participants of the experiments may be aware of such benchmarks and the target molecules used therein. Second, using exactly the same target molecule for one individual and one collective design experiment may be questionable, as participants of the first experiment may be in contact with participants of the second and could reveal the identity of the target molecules ahead of time. Third, as the target molecules of such benchmarks are approved drugs, they may be known by participants. To overcome such problems while ensuring the validity of the comparison with the benchmarks, 10 complexity-equivalent molecules were selected from ChEMBL database[48–52] between compounds that didn't reach clinical phases. In this way, meaning selecting real (i.e. non-virtual) molecules typical of the biologically active pre-clinical space, the necessity of choosing a pharmacology-relevant molecule was balanced with that of choosing a relatively unknown one.

To ensure the complexity equivalence between the chosen molecules and those used in original benchmarks, the following parameters were set to be the same: number of heavy atoms, number of aliphatic and aromatic rings, molecular fingerprints cardinality (i.e. the number of bits with a non-zero count in the molecular fingerprints) and the number of molecular fingerprints (i.e. the sum of all the individual fingerprints count). In this way, both a size- and complexity-equivalence were warranted. Target molecule complexity level is defined on the basis of fingerprints cardinality.

The original benchmark molecules and complexity-equivalent ones are shown in Table 1.

Complexity level	**Complexity features**	**Benchmark target molecule**	**Individual experiments target molecule**	**Collective experiments target molecule**
L1	**# heavy atoms**: 17 **# aliphatic rings**: 0 **# aromatic rings**: 1 **cardinality**: 33 **# fingerprints**: 45	 Albuterol	 T8 (CHEMBL460262)	 T9 (CHEMBL1159712)
L2	**# heavy atoms**: 26 **# aliphatic rings**: 0 **# aromatic rings**: 3 **cardinality**: 41 **# fingerprints**: 71	 Celecoxxib	 T13 (CHEMBL1566732)	 T32 (CHEMBL461573)

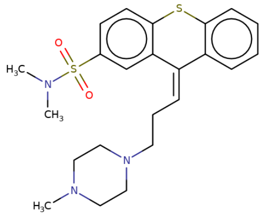
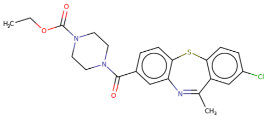
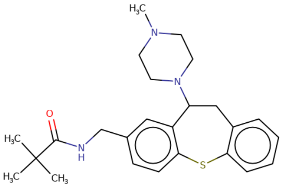
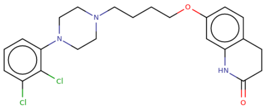
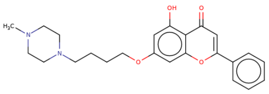
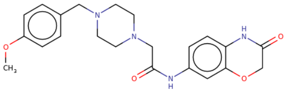
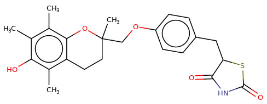
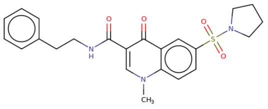
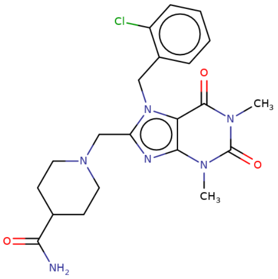
Complexity level	Complexity features	Benchmark target molecule	Individual experiments target molecule	Collective experiments target molecule
L3	**# heavy atoms**: 30 **# aliphatic rings**: 2 **# aromatic rings**: 2 **cardinality**: 51 **# fingerprints**: 85	 Thiothixene	 T15 (CHEMBL1352527)	 T14 (CHEMBL1259158)
L4	**# heavy atoms**: 30 **# aliphatic rings**: 2 **# aromatic rings**: 2 **cardinality**: 53 **# fingerprints**: 87	 Aripiprazole	 T19 (CHEMBL370628)	 T20 (CHEMBL554907)
L5	**# heavy atoms**: 31 **# aliphatic rings**: 2 **# aromatic rings**: 2 **cardinality**: 54 **# fingerprints**: 86	 Troglitazone	 T45 (CHEMBL2098358)	 T44 (CHEMBL1529981)

Table 1: Target molecules of the selected benchmarks and their corresponding complexity-equivalent target molecules used in this case study. For each complexity level, the common complexity features of the target molecules are reported. "Cardinality" is the number of bits with a non-zero count in the fingerprints of target molecules, while "# fingerprints" is the sum of all individual counts.

3.3 Molecular score

Every molecule designed in the system by participants was associated to a single-value molecular score. In all the experiments this score corresponded to the Tanimoto's similarity[53] towards its target molecule, linearly normalized in the 0-1000 range. The similarity was calculated using 1024-hashed, count-based, diameter-4, extended connectivity fingerprints (i.e. ECFC4_1024[54]) as implemented in CDK[55–58] (version 1.5.13). It has to be noted that such information was not shared with participants. The only two things they knew about the molecular score were its range and the fact that the higher the score, the closer the target molecule. The same molecular score but not normalized in the 0-1000 range was used for *de novo* design benchmarks comparison.

3.4 Experiment data, scoring and analysis

Each molecule created in the system may have been drawn starting from scratch or from another molecule already in the system. For each created molecule, the following information was stored *inter alia*: its structure, its score, its creator, its date and time of creation and the molecule from which it derived (if any). With this information it was possible to calculate different parameters to do a complete analysis of the experiments.

- **Maximum score reached**. The principal parameter used for the analysis is the maximum score reached in an experiment, represented by the top-1 molecular score calculated as explained in section 3.3. The maximum score reached is a measure of the efficacy achieved in an experiment.
- **Number of generated molecules**. An interesting parameter for evaluating the efficiency reached in experiments is the number of generated molecules. This corresponds to the number of unique molecules that are generated (and hence tested) to reach the final results. Uniqueness of molecules is calculated on basis of InChIKey, the hashed code derived from the standard InChI,[59] the IUPAC International Chemical Identifier.
- **Time played.** Another interesting parameter to evaluate the efficiency achieved in experiments is the time played, that is the total time spent by participants in designing molecules. Time played is computed considering the sum of the time frames between all the molecules designed by a participant. To avoid accounting for idle times, frames greater than one minute were omitted.
- **Scaffold/molecule ratio**. It is a parameter that can give information about how focused the molecular search is. This is the ratio between the number of unique molecules and unique scaffolds generated during one experiment. Scaffolds were defined according to Murko's definition[60] as calculated by RDKit.[d]
- **Number of molecule evolution steps**. Molecule design activity carried out by participants is divided into design sessions that correspond to different periods where molecules are designed. A design session includes all the molecules that are generated starting from scratch or from a certain molecule already in the system. The number of molecule evolution steps corresponds to the number of different design sessions needed for a certain molecule to be created. This is a particularly important and useful parameter for eventually found target molecules.
- **Collaboration degree**. It is defined as the percentage of experiment participants that are

[d] RDKit: Open-source cheminformatics. http:// www.rdkit.org.

involved in the creation of a certain molecule. It is a particularly important and useful parameter for eventually found target molecules.

- **Leader changes**. It is the number of times a new leader was recorded during an experiment, representing the events when a new participant overtakes the current highest score and search front.

3.5 Comparison with automatic *de novo* designers

In order to compare molecule design driven by human intelligence with that guided by artificial intelligence (i.e. *de novo* designers), this case study was oriented on GuacaMol,[43] a recently published benchmark suite for *de novo* molecular design. There, two types of benchmarks are proposed: the distribution-learning benchmarks that assess the capacity of a method to reproduce the distribution of a certain molecule set and the goal-directed benchmarks that evaluate the ability to generate individual molecules with predefined features (i.e. molecules can be scored individually). The use of GuacaMol goal-directed benchmarks allows to compare the molecular search strategy of humans with that of some recent *de novo* designers considered state-of-the-art in the field. These systems represent a variety of searching methods as: genetic algorithms (GA),[61] Long-Short Term Memory recurrent neural networks (LSTM)[62] and Monte Carlo Tree Search (MCTS)[63] applied to two molecular representations: graph-based and SMILES-based.[64,65] In total the following five baseline models are considered in GuacaMol for goal-directed benchmark: *smiles_ga*,[66] *graph_ga*,[42] *graph_mcts*,[42] *smiles_lstm*[38] and *best_of_dataset*. Where: the first four are named after the used molecular representation and the used searching algorithm type, while the fifth is a database virtual screening. This last represents the minimal score and only *de novo* search strategies that score higher have an advantage over simple virtual screening.

The first five goal-directed benchmarks of GuacaMol were selected, consisting of the three rediscovery and the two similarity benchmarks reported in Table 2.

Benchmark name	**Benchmark type**	**Scoring function**	**Scoring**
Celecoxxib rediscovery	Rediscovery	sim(Celecoxxib, ECFC4)	Top-1
Troglitazone rediscovery	Rediscovery	sim(Troglitazone, ECFC4)	Top-1
Thiothixene rediscovery	Rediscovery	sim(Thiothixene, ECFC4)	Top-1

Aripiprazole similarity	Similarity	Thresholded(0.75)
		sim(Aripiprazole, ECFC4)
			Top-1, top-10, top-100

Albuterol similarity	Similarity	Thresholded(0.75)
		sim(Albuterol, ECFC4)
			Top-1, top-10, top-100

Table 2: Benchmarks selected from GuacaMol.[43] "Scoring" refers to the number of top molecules considered in the score calculation.

The aim of a rediscovery benchmark is to evaluate the rediscovery (i.e. re-design) of a single target molecule of interest, while that of a similarity benchmark is to evaluate the generation of many molecules that are closely related to a single target molecule. The scoring function used in the first case is the Tanimoto's similarity[53] to the target molecule calculated using ECFC4 fingerprints, while the second one uses the same scoring function adjusted with a 0.75-threshold modifier. As described in the original publication,[43] such modifier assigns a full score (i.e. 1.0) to values above a given threshold *t* (in this cases 0.75) while values smaller than *t* decrease linearly to zero. Finally, rediscovery benchmarks base their score on the top-1 molecule generated during the design, while similarity ones on the top-1, top-10, top-100 molecules and their average.

4 Results and discussion

4.1 Participation

After the scientific community was called to engage in the case study as described in section 3.1, the participation results reported in Table 3 were obtained. A total of 118 participants completed the sign up process; 91 of them accessed the *Sandbox*, where they could learn the basics of the application; 71 completed the *Sandbox* requirements and were invited to the experiments; 46 took finally part in the experiments and 31 of them resulted to be very active, drawing more than 100 molecules each.

Event	**Participants**
Sign up process completion	118
Sandbox access	91
Sandbox completion	71
Participation in challenges	46
High activity in challenges (> 100 drawn molecules)	31

Table 3: Participation results.

46 participants of the initial 118 who signed up (i.e. 39%) engaged in the experiments but 71 out of 91 (78%) who accessed the *Sandbox* could correctly complete its requirements. This means that loss of participants in relation to the possible difficulty of using the application (i.e. 20) represents only 28% of all drop outs, highlighting the ease of participating in the case study. The choice to demand the completion of the *Sandbox* requirements before letting the participants access the challenges allowed them to learn the basics of the application and practice with it without tampering with the data generated in the experiments.

Each of the 71 participants that completed the *Sandbox* requirements was invited to one individual and one collective experiment. The average invitation was 12 participants per experiment while the average engagement (people who draw at least 1 molecule) was 7.

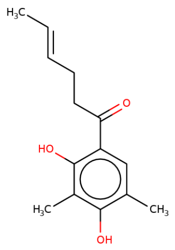
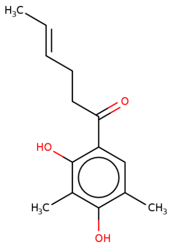
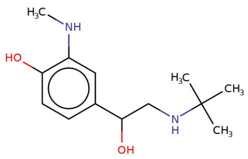
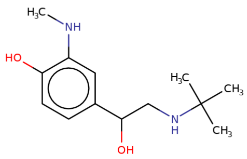
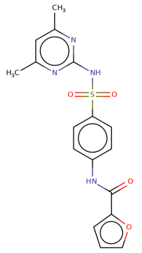
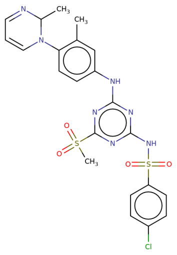
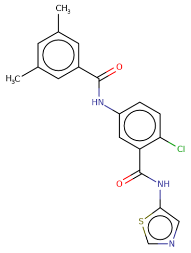
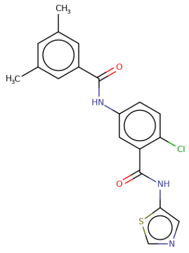
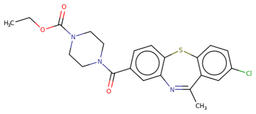
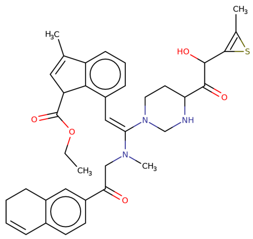
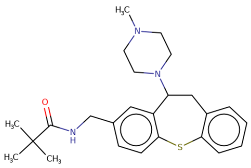
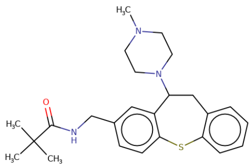
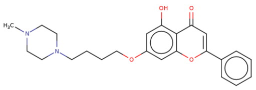
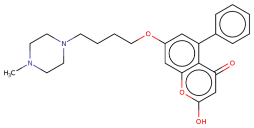
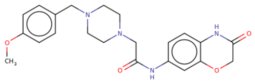
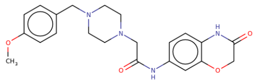
To achieve the highest number of participants, it was opted to keep participant profiling as basic as possible. The participants were asked for the following information: their full name, their e-mail address, and if they studied chemistry/biotechnology/biology or a related discipline so that they feel comfortable in sketching molecular structures (condition for which participants were denoted here as “skilled participants”). Only eight non-skilled participants completed the sign up process, but none of them completed the *Sandbox* necessary to participate in the experiments. In order to achieve similar levels of human knowledge in the individual and collective settings, participants were invited to both experiment types. 83% of people who participated in the collective experiments also participated in the individual ones (see *Supporting Information).* Finally, it’s worth to mention that no single participant in collective experiments overperformed compared to the others so that the hypothesis that a single participant drove the full collective experiment could be excluded.

The case study successfully recruited dozens of active participants which allowed an acceptable analysis of the observed tendencies and behaviors. Although achieving hundreds or even thousands of active participants would certainly be favorable to obtain more statistically significant results, such ideal situation is very difficult to achieve. A first difficulty is getting into contact and motivating enough skilled participants to enroll in and drive the experiments. In this respect even for very successful scientific games as FoldIt, that achieved thousands of sign-ups, most of the puzzles, comparable to our experiments, were basically led by less than 10 people per puzzle (5 people being the median and 6 the mean)[4] . Such few participants were those who improved the experiment score. Another difficulty, triggered to enable collective design dynamics, was that in our case people should participate synchronously in the experiments during at most 2 weeks.

Communication of these first results could, as in case of other scientific on-line games, raise the participation number in future challenges to further support the statistical significance.

4.2 Finding the target molecules

In total, 10 different experiments were conducted to assess human search strategy in chemical space exploration: five individual and five collective ones. Results are reported in Table 4.

Complexity level	Individual design		Collective design	
	Target molecule	Best molecule achieved	Target molecule	Best molecule achieved
L1	 T8	 Score = 1000	 T9	 Score = 1000
L2	 T13	 Score = 722	 T32	 Score = 1000
L3	 T15	 Score = 605	 T14	 Score = 1000
L4	 T19	 Score = 931	 T20	 Score = 1000

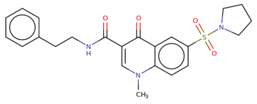
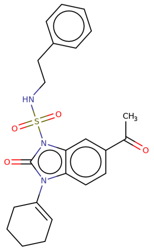
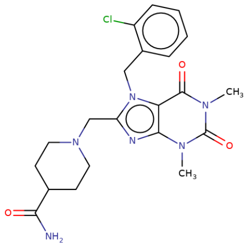
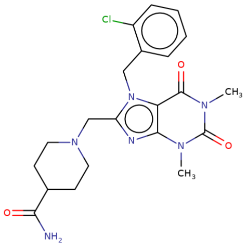
Complexity level	Individual design		Collective design	
	Target molecule	Best molecule achieved	Target molecule	Best molecule achieved
L5	 T45	 Score = 802	 T44	 Score = 1000

Table 4: Target and best (i.e. most similar) molecules designed by participants in individual and collective experiments.

The first very important result is that in several experiments participants were able to find the target molecule (i.e. score = 1000), that is one specific, predefined molecule among the almost infinite possibilities in the huge chemical space. As far as we know, this is the first time that such a study, quantifying molecule search strategy of humans, is conducted. This result is particularly important considering the following circumstances:

1. Participants searched the chemical space from scratch by drawing molecules starting from a simple carbon atom.
2. As molecules are drawn and manipulated on an atom/bond level, participants had absolute freedom to potentially reach any organic drug-like molecule of the chemical space.
3. Participants searched the chemical space simply by following a single-value molecular score indicating how close they were to the target molecule. They didn't receive any additional hint or information and had to build their own logic behind it.

Target molecules of five different complexity levels were searched. In individual design experiments, participants could only find the most simple target molecule (i.e. T8). Anyway, in the cases of the two most complex targets (i.e. T19 & T45), they got close and reached scores of 931 and 802, corresponding to a Tanimoto's molecular similarity of 0.931 and 0.802, respectively. In contrast, in collective design experiments participants could find the target molecule in all the cases.

The scoring function that should be followed in a real drug design program aiming to reach lead compounds would certainly be more complex than the simple similarity function used in such experiments. Indeed it should consider not only the compounds capacity of interacting with the biological target of interest but also their pharmacokinetics (i.e. ADME) and toxicity (T) profile, elements that can be predicted *in silico* by machine learning models. The choice of using a similarity function in this case study was dictated by two main reasons: i) it is a surrogate for machine learning models and if a *de novo* molecular generator doesn't work using similarity functions, probably it will have difficulties in working with more complex functions (this is also why similarity functions are used as basic functions in *de novo* design benchmarking). ii) the molecular generation results obtained using a similarity function are easily interpretable and the achievement of the predefined target molecule, maximizing the molecular scoring, can be unequivocally determined. Using a similarity function is therefore a useful first step to take before searching more complex scenarios which results have an undoubted intrinsic value.

4.3 Individual vs collective molecule design

Experiment results are reported in Table 5.

Target complex. level	**Individual design**						**Collective design**					
	Target mol.	**Time played**	**Generated unique molecules**	**Scaffold/ molecule ratio**	**Leader changes**	**Max score[e]**	**Target mol.**	**Time played**	**Generated unique molecules**	**Scaffold/ molecule ratio**	**Leader changes**	**Max score[e]**
L1	T8	6H 24m	2,402	0.184	6	**1,000**	T9	2H 45m	1,343	0.186	6	**1,000**
L2	T13	16H 13m	6,821	0.429	11	722	T32	7H 10m	2,936	0.266	11	**1,000**
L3	T15	9H 53m	4,544	0.325	9	605	T14	9H 19m	3,708	0.246	9	**1,000**
L4	T19	27H 42m	11,660	0.384	7	931	T20	7H 34m	2,856	0.384	15	**1,000**
L5	T45	11H 31m	5,971	0.381	3	802	T44	19H 40m	7,842	0.404	13	**1,000**

Table 5: Results obtained by participants in the individual & collective search for specific, predefined target molecules in the chemical space. Target molecules are classified by complexity level. The number of generated unique molecules is reported together with the scaffold/molecule ratio. Leader changes represent the number of times a new leader was recorded during an experiment. The max score is the highest score obtained in an experiment (max = 1,000).

The following observations can be made on the basis of the results:

1. **Collective design seems more efficacious than individual design.** While in the five individual design experiments the target molecule was found only in the simplest case, all the five collective design experiments were successful. This suggests a higher efficacy of collective molecule design in respect to individual one.

2. **Collective design seems more efficient than individual design.** Collective design succeeded in finding the target molecule not only by generating (and hence testing) less molecules but also by needing less playing time. There is just one case where the collective design generated more molecules and took more playing time than the individual one: the experiment targeting the most complex target molecule (i.e. complexity level L5). Nevertheless, as the individual search could not find the target molecule, it cannot be concluded that in this case individual design was more efficient.

3. **Collective search is at least as broad as the individual one.** One concern about collective design may be that, given a certain number of molecules, it generates less scaffolds in respect to the individual design. This may happen as at any moment in time all participants may center

[e] This is the molecular score visible by the participants in the application. It is different from the scores calculated for the *de novo* design benchmarks comparison.

their search around the best molecule (or currently few best molecules) so that fewer scaffolds are generated. This hypothesis seems to be incorrect as it only holds up in two out of five cases, which can be seen on basis of the scaffold-molecule ratio reported in Table 5.

4. **Designing complexity.** Interestingly, the number of molecules needed by collective design to reach the target molecule does not correlate with its computationally estimated complexity. Similarly, in case of individual experiments the maximum score achieved does not inversely correlate with the target complexity metrics as it could be expected. This may indicate that the designing complexity experienced by humans differs from the one computationally defined.
5. **Collaboration.** The collaboration degree of target molecules achieved in collective design experiments ranges from 50% to 100%. In two of the four experiments where collective design was more efficacious than individual design, more leader changes are observed. Interestingly, the difference is particularly large in case of the two most complex targets (i.e. 15 vs 7 and 13 vs 3 for collective vs individual experiments with target molecule complexity level L4 and L5, respectively). It can be hypothesized that leader changes in collective design is beneficial for reaching the objective.

The evolution steps of the target molecule (defined in *Methods*) achieved in the five successful collective design experiments ranged from 9 to 29 while their collaboration degree ranged from 50% to 100%. As the possibility to collaborate is the only setting difference between the individual and collective experiments, the high collaboration degree in the creation of the target molecules may be the cause for the higher efficacy achieved in the collective experiments.

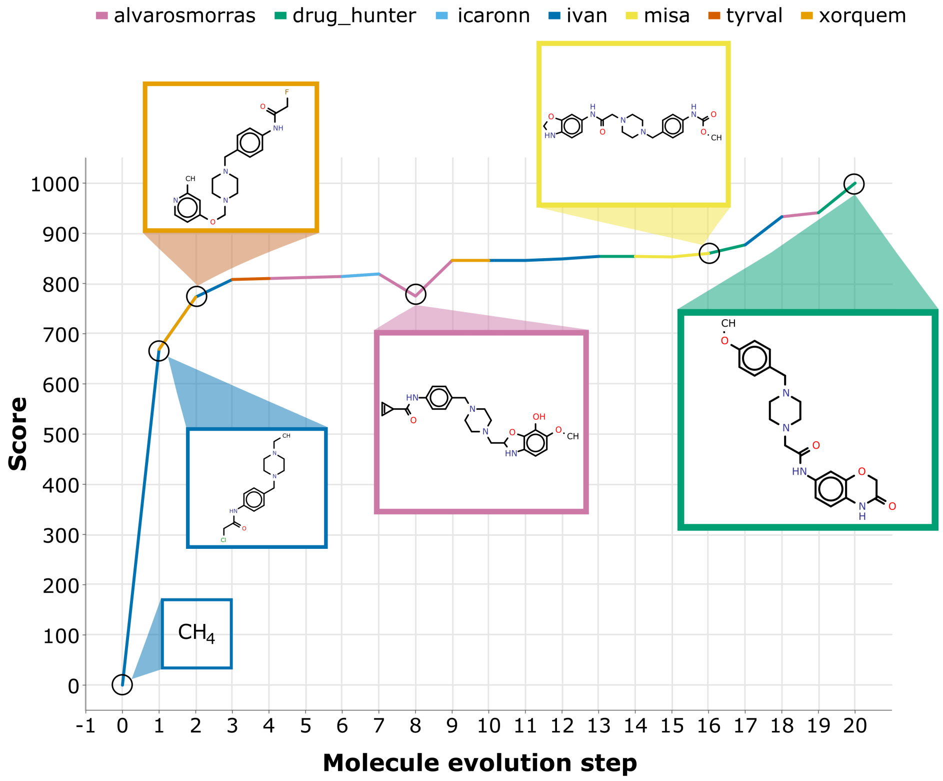


Figure 1: Genesis of target molecule T20. The target molecule is created (i.e. rediscovered) in 20 evolution steps through the collective design efforts of seven *out of* eight *participants of this experiment. The individual contributions to the target molecule creation are represented by different colors. Some intermediate generated molecules are also shown.*

To illustrate such features, the genesis of target molecule T20 is reported in Figure 1.Target molecule T20 was generated in 20 evolution steps through the collective work of seven out of the eight participants of this experiment. While the general trend of molecule evolution is positive, meaning the score of the resulting molecule in each design session is higher than the starting molecule, there are evolution steps in the genesis of target T20 where the score remains equal (steps 10, 11 and 14) or even decreases (steps 8 and 15). The transit through molecules with scores lower than the experiment maximum may represent the exit mechanism from local maxima.

To better understand the differences between individual and collective design, experiments related to complexity-level-L4 target molecules (i.e. T19 & T20) are compared.

The top-score achieved by each user along the whole molecule design activity of L4-complexity targets experiments is represented in Figure 3.

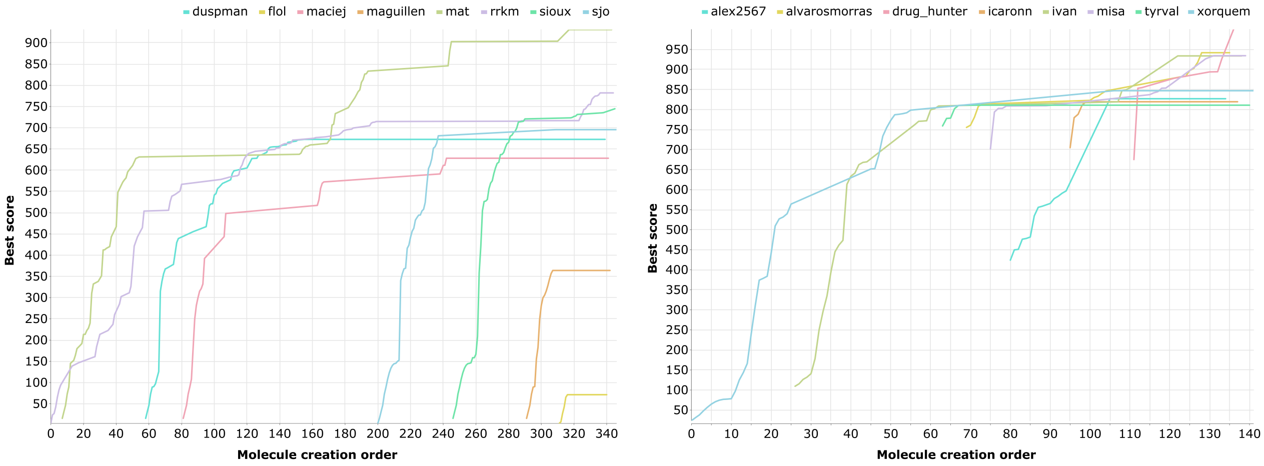


Figure 2: Molecule best scores (y-axis) achieved by participants during individual (left) and collective (right) design experiments of L4-complexity-level target molecule. Molecule creation order (x-axis) is the order in which the best molecules (i.e. user-based top-1 molecules) are generated.

A first consideration is that it seems easier for participants to rise the molecule score from 0 to around 550, than from around 550 to 1000. This is an expected behavior. On one side this may be due to the fact that similarity may rise quickly when some common functional groups are initially added to the structure. On another side, however, this behavior may also reflect a general feature of the chemical space search: it is more difficult to design an optimal molecule (i.e. max score) than a sub-optimal one.

Two main differences emerge from the comparison of the two plots reported in Figure 2:

- While in the individual design experiment all the participants started the design activity from molecules with a score close to 0, in the collective design one all but the first started exploring the chemical space from already designed molecules with higher scores.
- While the number of leader changes in the individual challenge is limited (i.e. 7), in the collective challenge it is significantly higher (i.e. 15) as everybody can start from the highest scoring molecule.

To understand the structural diversity of the molecules generated during a design experiment, their distribution in the chemical space can be examined. For such a purpose, molecules are first characterized using the same descriptors with which the molecular score was calculated (i.e. 1024-hashed ECFC4 fingerprints) and then plotted in Figure 3 using t-SNE (i.e. t-distributed stochastic neighbor embedding).[67]

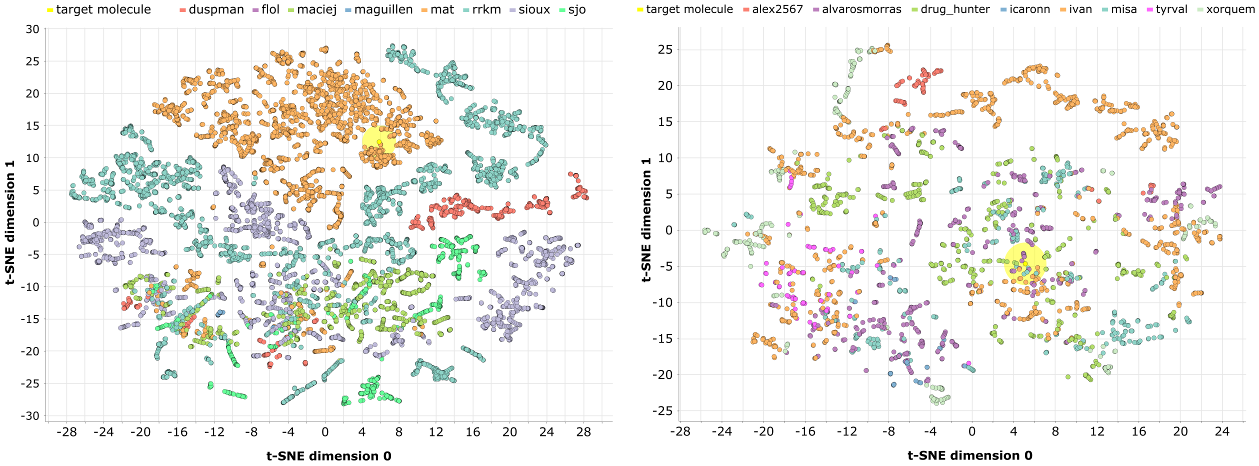


Figure 3: Chemical space explored by each participant during an individual (left) and a collective (right) design experiment. Molecules are described by 1024-hashed ECFC4 fingerprints and represented using a t-SNE visualization. The molecules generated by each participant are represented by a different color. The target molecule is represented by a yellow point highlighted by a halo.

The following observations can be made about the chemical space plots:

- While in the individual design experiment it seems that specific participants explored specific, focused parts of the chemical space, in the collective design one the molecules generated by each user are more spread in the chemical space.
- In the individual design experiment only one participant came close to the target molecule, while in the collective design one at least four of them.

4.4 Comparison with automatic *de novo* designers

As described in section 3.5 this study was designed to compare the search strategy of humans with automatic *de novo* designers. For such a purpose a recently published *de novo* design benchmark was chosen that includes results from different automatic methods. Its usage allows also to dissipate any possible doubt that could have arisen if we would have used internal automatic *de novo* systems for comparison.

The results of both individual and collective human design activity for the five selected benchmarks are reported in Figure 4 and Table 6 together with those of the state-of-the-art *in silico* methods published in the original benchmark article.

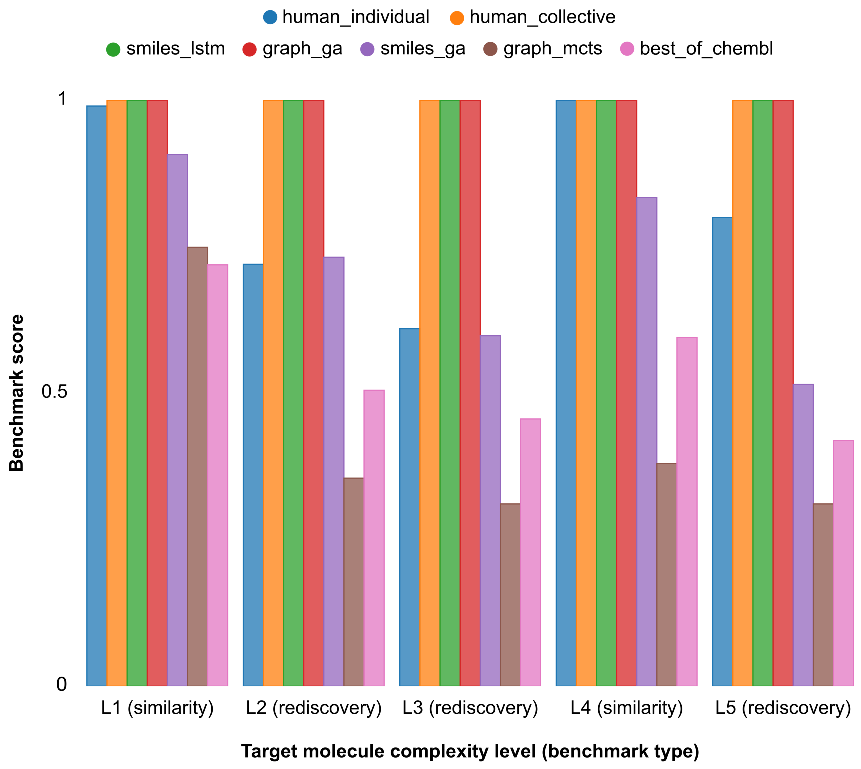


Figure 4: Comparison of human individual and collective design experiments with in silico de novo designers reported in the GuacaMol publication.[43] *Benchmark scores are explained in section 3.5.*

		Final score Top-1 score (Top-10 score) (Top-100 score)						
Complexity level	**Benchmark type**	**Human individual**	**Human collective**	**smiles_lstm**	**graph_ga**	**smiles_ga**	**graph_mcts**	**best_of_chembl**
L2	Rediscovery	**0.72** 0.72	**1.0** 1.0	**1.0** 1.0	**1.0** 1.0	**0.732** 0.732	**0.355** 0.355	**0.505** 0.505
L3	Rediscovery	**0.61** 0.61	**1.0** 1.0	**1.0** 1.0	**1.0** 1.0	**0.598** 0.598	**0.311** 0.311	**0.456** 0.456
L5	Rediscovery	**0.80** 0.80	**1.0** 1.0	**1.0** 1.0	**1.0** 1.0	**0.515** 0.515	**0.311** 0.311	**0.419** 0.419
L1	Similarity	**0.99** 1.0 1.0 0.96	**1.0** 1.0 1.0 1.0	**1.0** 1.0 1.0 1.0	**1.0** 1.0 1.0 1.0	**0.907** 1.0 1.0 0.72	**0.749** 0.80 0.758 0.689	**0.719** 0.765 0.726 0.664
L4	Similarity	**1.0** 1.0 1.0 1.0	**1.0** 1.0 1.0 1.0	**1.0** 1.0 1.0 1.0	**1.0** 1.0 1.0 1.0	**0.834** 0.856 0.838 0.807	**0.380** 0.428 0.376 0.335	**0.595** 0.609 0.601 0.576

Table 6: Comparison of human individual and collective design experiments with automatic de novo designers reported in the GuacaMol publication.[43] *Benchmark scores are explained in section 3.5. The final score is equivalent to the top-1 score in rediscovery benchmarks and to the average of top-1, top-10 and top-100 scores in the similarity ones.*

Human collective design performed optimally along all the five tested benchmarks. This is also the case for the two best automatic systems (i.e. smiles_lstm[38] and graph_ga[42]). Human individual design performed more poorly than collective design but still fairly well. Actually, in case of the similarity benchmarks, it achieved almost the optimal scores (i.e. 1.0 and 0.99 in experiments with targets of L4 and L1 complexity level, respectively), while in the case of the rediscovery benchmarks it performed worse than the two best *in silico* systems, but better than two out of the three other approaches.

In the cases where the benchmark maximum score of 1.0 is not reached, the relation between the complexity of the target molecules and the achieved efficacy is analyzed. Here, efficacy is determined by how close the final achieved score is to the maximum (i.e. 1.0). Interestingly, automatic methods correlate inversely with the estimated complexity levels of the target molecules while this is not true for human individual design. More specifically, this occurs in rediscovery benchmarks (L2, L3 and L5) where smiles_ga =0.732, 0.598, 0.515, graph_mcts =0.355, 0.311, 0.311 and human_individual = 0.72, 0.61, 0.80, respectively . This also occurs in similarity benchmarks (L1 and L4) where smiles_ga = 0.907, 0.834; graph_mcts = 0.749, 0.380; human_individual 0.99, 1.0, respectively. While for the automatic methods the molecular design difficulty seems to correlate with the computationally estimated complexity of target molecules, this does not hold up for human design activity.

4.5 Human vs machine learning pace

A possible measure for the learning pace of the search strategy is the number of times the molecular scoring function has been accessed for finding a particular target molecule. The higher the number, the slower the learning pace. In case of human-driven *de novo* design described herein, this is the number of moves carried out by participants to reach the target molecule. This corresponds to all the (non-unique) molecules generated in the experiments. This number is larger than the number of generated unique molecules reported in Table 5, because it also considers repetitions. In other words, if the same molecule has been drawn five times, it will count as five scoring function calls.

The number of scoring function calls carried out by individual and collective human intelligence are reported in Table 7 together with those of Long-Short Term Memory recurrent neural networks (lstm_smiles)[38], reported[f] in the GuacaMol[43] publication. Human individual design results are only reported for the experiment where participants reached the target molecule.

[f] This is the only method for which a reliable number of scoring function calls is reported in the original publication (private communication with authors).

	Number of scoring function calls to reach the target molecule		
Complexity level	**lstm_smiles**	**Human individual**	**Human collective**
L1	132,838	3,614	1,956
L2	132,846	-[g]	4,271
L3	138,209	-[g]	5,404
L4	139,221	-[g]	4,591
L5	140,339	-[g]	12,118

Table 7: Number of scoring function calls needed to reach the target molecules of five *different complexity levels. Human individual design results are only reported for the experiment where participants reached the target molecule.*

It can be seen that the number of scoring function calls carried out by humans (in both the individual and collective design mode) are more than one order of magnitude lower than those of the artificial neural network. These results suggest that humans have a larger learning pace in respect to the considered AI method. The learning pace is related with the efficiency.

Interestingly, while the number of scoring function calls needed by artificial intelligence (i.e. lstm_smiles) to reach the target molecule correlates with its complexity level, this does not occur with human intelligence. This observation was also done for efficacy as described above.

This trend should be taken with caution as other AI methods could work differently.

5 Conclusions

In the last decade individual and collective human intelligence were used in combination with computer algorithms to solve complex scientific problems. These are problems with many degrees of freedom where computational algorithms alone struggle to find the best solution. This approach was successfully used in different research fields as comparative genomics, structural biology, macromolecular crystallography and RNA design. Here we described an attempt to use a similar approach in small-molecule drug design. More specifically we assessed the human search strategy in chemical space exploration problems where specific, predetermined molecules had to be found between the almost infinite possibilities. Finally, results were compared to those obtained by different automatic *de novo* designers assessed in a recently published benchmark suite. This allows to have a first direct comparison between human and artificial intelligence in *de novo* drug design.

The here explained case study focused on the usage of a similarity function as design scoring. Although this is certainly a simplification in respect to a drug discovery scenario where more complex multi-objective scoring functions should be used, the molecular similarity is a surrogate for machine learning models and have the advantage of producing easily interpretable results where the achievement of predefined target molecules can be unequivocally determined. In this respect, this study should be regarded as a first necessary step towards the usage of the same approach with more complex scoring

[g] Target molecule not reached.

functions.

From the results, the following conclusions can be drawn:

1. The search strategy linked to human intelligence can be successfully used in chemical space exploration *in silico*. Indeed, it is able to find unique, predefined target molecules, having a molecular complexity equivalent to that of approved drugs, between the huge amount of possibilities. This supports the usage of human search capability coupled to *in silico* molecule evaluation systems in drug design.
2. Collective human molecular design seems to be both more efficacious and more efficient than individual molecular design. This supports the development of collaborative drug design tools that allow to create synergies between different players of this field and reach better drugs.
3. Compared to artificial intelligence systems, the search efficacy of human collective intelligence seems to be at least as good as the best artificial intelligence approaches. In contrast, human individual intelligence ranks average. Considering the search efficiency, these first results suggest that human intelligence may have a higher learning pace than artificial intelligence. This observation should be endorsed by additional experiments including a higher number of AI systems.

Additionally, some results may suggest that human intelligence perceives molecular complexity differently than artificial intelligence but also in this case more experiments will be needed to confirm such finding. If confirmed, this would support a combined use of the two intelligences in order to reach better drugs. In our group we are currently working on two main topics: the extension of the current study with more complex, multi-objective scoring functions; and the implementation of a hybrid *de novo* designer where human and artificial intelligences are integrated in a unique system.

6 Acknowledgment

The authors warmly thank all the people who participated in this case study (see Tables S1 and S2 in *Supporting Information*). Due to their help, a first understanding of human intelligence in exploring the chemical space has been achieved. We are very grateful to Prof. Miquel Pons for reading an early version of this manuscript and giving helpful feedback.

7 Bibliography

(1) Curtis, V. Online Citizen Science Games: Opportunities for the Biological Sciences. *Appl. Transl. Genomics* **2014**, *3*, 90–94.

(2) Treuille, A.; Das, R. Scientific Rigor through Videogames. *Trends Biochem. Sci.* **2014**, *39*, 507–509.

(3) Rowles, T. A. Power to the People: Does Eterna Signal the Arrival of a New Wave of Crowd-Sourced Projects? *BMC Biochem.* **2013**, *14*, 1.

(4) Cooper, S.; Khatib, F.; Treuille, A.; Barbero, J.; Lee, J.; Beenen, M.; Leaver-Fay, A.; Baker, D.; Popović, Z.; Players, F. Predicting Protein Structures with a Multiplayer Online Game. *Nature* **2010**, *466*, 756–760.

(5) Khatib, F.; Dimaio, F.; Cooper, S.; Kazmierczyk, M.; Gilski, M.; Krzywda, S.; Zabranska, H.; Pichova, I.; Thompson, J.; Popović, Z.; Jaskolski, M.; Baker, D. Crystal Structure of a Monomeric Retroviral Protease Solved by Protein Folding Game Players. *Nat. Struct. Mol. Biol.* **2010**, *18*, 1175–1177.

(6) Khatib, F.; Cooper, S.; Tyka, M. D.; Xu, K.; Makedon, I.; Popovic, Z.; Baker, D.; Players, F. Algorithm Discovery by Protein Folding Game Players. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 18949–18953.

(7) Horowitz, S.; Koepnick, B.; Martin, R.; Tymieniecki, A.; Winburn, A. A.; Cooper, S.; Flatten, J.; Rogawski, D. S.; Koropatkin, N. M.; Hailu, T. T.; Jain, N.; Koldewey, P.; Ahlstrom, L. S.; Chapman, M. R.; Sikkema, A. P.; Skiba, M. A.; Maloney, F. P.; Beinlich, F. R. M.; Popovic, Z.; Baker, D.; Khatib, F.; Bardwell, J. C. A. Determining Crystal Structures through Crowdsourcing and Coursework. *Nat. Commun.* **2016**, *7*, 1–9.

(8) Koepnick, B.; Flatten, J.; Husain, T.; Ford, A.; Silva, D. A.; Bick, M. J.; Bauer, A.; Liu, G.; Ishida, Y.; Boykov, A.; Estep, R. D.; Kleinfelter, S.; Nørgård-Solano, T.; Wei, L.; Players, F.; Montelione, G. T.; DiMaio, F.; Popović, Z.; Khatib, F.; Cooper, S.; Baker, D. De Novo Protein Design by Citizen Scientists. *Nature* **2019**, *570*, 390–394.

(9) Kawrykow, A.; Roumanis, G.; Kam, A.; Kwak, D.; Leung, C.; Wu, C.; Zarour, E.; Sarmenta, L.; Blanchette, M.; Waldispühl, J. Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment. *PLoS One* **2012**, *7*.

(10) Jorda, J.; Sawaya, M. R.; Yeates, T. O. CrowdPhase: Crowdsourcing the Phase Problem. *Acta Crystallogr. D. Biol. Crystallogr.* **2014**, *70*, 1538–1548.

(11) Jorda, J.; Sawaya, M. R.; Yeates, T. O. Progress in Low-Resolution Ab Initio Phasing with CrowdPhase. *Acta Crystallogr. Sect. D Struct. Biol.* **2016**, *72*, 446–453.

(12) Levieux, G.; Tiger, G.; Mader, S.; Zagury, J. F.; Natkin, S.; Montes, M. Udock, the Interactive Docking Entertainment System. *Faraday Discuss.* **2014**, *169*, 425–441.

(13) Levieux, G.; Montes, M. Towards Real-Time Interactive Visulaization Modes of Molecular Surfaces: Examples with Udock. In *1st International Workshop on Virtual and Augmented Reality for Molecular Science (VARMS)*; Arles, France, 2015.

(14) Lee, J.; Kladwang, W.; Lee, M.; Cantu, D.; Azizyan, M.; Kim, H.; Limpaecher, A.; Gaikwad, S.; Yoon, S.; Treuille, A.; Das, R. RNA Design Rules from a Massive Open Laboratory. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 2122–2127.

(15) Koodli, R. V.; Keep, B.; Coppess, K. R.; Portela, F.; Das, R. EternaBrain: Automated Rna Design through Move Sets and Strategies from an Internet-Scale RNA Videogame. *PLoS Comput. Biol.* **2019**, *15*, 1–22.

(16) Bohacek, R. S.; Mcmartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design : A Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.

(17) Fink, T.; Bruggesser, H.; Reymond, J.-L. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons. *Angew. Chem. Int. Ed. Engl.* **2005**, *44*, 1504–1508.

(18) Drew, K. L. M.; Baiman, H.; Khwaounjoo, P.; Yu, B.; Reynisson, J. Size Estimation of Chemical Space: How Big Is It? *J. Pharm. Pharmacol.* **2012**, *64*, 490–495.

(19) Reymond, J.-L.; van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical Space as a Source for New Drugs. *Medchemcomm* **2010**, *1*, 30.

(20) Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discove. *J. Chem. Inf. Model.* **2007**, *47*, 342–353.

(21) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48*, 722–730.

(22) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput. Aided. Mol. Des.* **2013**, *27*, 675–679.

(23) Ruddigkeit, L.; Deursen, R. Van; Blum, L. C.; Reymond, J. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

(24) Walters, W. P. Virtual Chemical Libraries. *J. Med. Chem.* **2019**, *62*, 1116–1124.

(25) Schneider, G. *De Novo Molecular Design*; Schneider, G., Ed.; Wiley-VCH, 2014.

(26) Hartenfeller, M.; Schneider, G. Enabling Future Drug Discovery by de Novo Design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 742–759.

(27) Böhm, H. J. The Computer Program LUDI: A New Method for the de Novo Design of Enzyme Inhibitors. *J. Comput. Aided. Mol. Des.* **1992**, *6*, 61–78.

(28) Gillet, V. J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P. SPROUT: Recent Developments in the de Novo Design of Molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 207–217.

(29) Wang, R.; Gao, Y.; Lai, L. LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design. *J. Mol. Model.* **2000**, *6*, 498–516.

(30) Brown, N.; McKay, B.; Gilardoni, F. A Graph-Based Genetic Algorithm and Its Application to the Multiobjective Evolution of Median Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1079–1087.

(31) Ertl, P.; Lewis, R. IADE: A System for Intelligent Automatic Design of Bioisosteric Analogs. *J. Comput. Aided. Mol. Des.* **2012**, *26*, 1207–1215.

(32) Nicolaou, C. A.; Apostolakis, J.; Pattichis, C. S. De Novo Drug Design Using Multiobjective Evolutionary Graphs. *J. Chem. Inf. Model.* **2009**, *49*, 295–307.

(33) Firth, N. C.; Atrash, B.; Brown, N.; Blagg, J. MOARF, an Integrated Workflow for Multiobjective Optimization: Implementation, Synthesis, and Biological Evaluation. *J. Chem. Inf. Model.* **2015**, *55*, 1169–1180.

(34) Yang, X.; Zhang, J.; Yoshizoe, K.; Terayama, K.; Tsuda, K. ChemTS: An Efficient Python Library for de Novo Molecular Generation. *Sci. Technol. Adv. Mater.* **2017**, *18*, 972–976.

(35) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning:Generative Models for Matter Engineering. *Science (80-.).* **2018**, *361*, 360–365.

(36) You, J.; Liu, B.; Ying, R.; Pande, V.; Leskovec, J. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. *Adv. Neural Inf. Process. Syst.* **2018**, *2018-Decem*, 6410–6421.

(37) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A.

Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.

(38) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120–131.

(39) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for de Novo Drug Design. *Sci. Adv.* **2018**, *4*, 1–15.

(40) Oboyle, N. M.; Campbell, C. M.; Hutchison, G. R. Computational Design and Selection of Optimal Organic Photovoltaic Materials. *J. Phys. Chem. C* **2011**, *115*, 16200–16210.

(41) Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-like Compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296–7303.

(42) Jensen, J. H. A Graph-Based Genetic Algorithm and Generative Model/Monte Carlo Tree Search for the Exploration of Chemical Space. *Chem. Sci.* **2019**, *10*, 3567–3572.

(43) Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. GuacaMol: Benchmarking Models for De Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 1096–1108.

(44) Hack, M. D.; Rassokhin, D. N.; Buyck, C.; Seierstad, M.; Skalkin, A.; Ten Holte, P.; Jones, T. K.; Mirzadegan, T.; Agrafiotis, D. K. Library Enhancement through the Wisdom of Crowds. *J. Chem. Inf. Model.* **2011**, *51*, 3275–3286.

(45) Sheridan, R. P.; Zorn, N.; Sherer, E. C.; Campeau, L. C.; Chang, C.; Cumming, J.; Maddess, M. L.; Nantermet, P. G.; Sinz, C. J.; O'Shea, P. D. Modeling a Crowdsourced Definition of Molecular Complexity. *J. Chem. Inf. Model.* **2014**, *54*, 1604–1616.

(46) Boobier, S.; Osbourn, A.; Mitchell, J. B. O. Can Human Experts Predict Solubility Better than Computers? *J. Cheminform.* **2017**, *9*, 1–14.

(47) Bush, J. T.; Pogány, P.; Pickett, S. D.; Barker, M.; Baxter, A.; Campos, S.; Cooper, A. W. J.; Hirst, D. J.; Inglis, G.; Nadin, A.; Patel, V. K.; Poole, D.; Pritchard, J.; Washio, Y.; White, G.; Green, D. A Turing Test for Molecular Generators. *J. Med. Chem.* **2020**, *63*, 11964–11971.

(48) Gaulton, A.; Bellis, L. J.; Bento, a P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100-7.

(49) Bento, a P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. a; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083-90.

(50) Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J. P. Activity , Assay and Target Data

Curation and Quality in the ChEMBL Database. *J. Comput. Aided. Mol. Des.* **2015**, *29*, 885–896.

(51) Gaulton, A.; Hersey, A.; Patr, A.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibri, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magari, P.; Overington, J. P.; Papadatos, G.; Smit, I. The ChEMBL Database in 2017. **2017**, 1–10.

(52) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47*, D930–D940.

(53) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform.* **2015**, *7*, 1–13.

(54) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(55) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.

(56) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2111–2120.

(57) May, J. W.; Steinbeck, C. Efficient Ring Perception for the Chemistry Development Kit. *J. Cheminform.* **2014**, *6*, 1–12.

(58) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; Torrance, G.; Evelo, C. T.; Guha, R.; Steinbeck, C. The Chemistry Development Kit (CDK) v2.0: Atom Typing, Depiction, Molecular Formulas, and Substructure Searching. *J. Cheminform.* **2017**, *9*, 1–19.

(59) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. *InChI, the IUPAC International Chemical Identifier*; Journal of Cheminformatics, 2015; Vol. 7.

(60) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs . 1 . Molecular Frameworks. *J. Med. Chem.* **1996**, *2623*, 2887–2893.

(61) Mitchell, M. *Introduction to Genetic Algorithms*; MIT Press, 1996, Ed.; 5th ed.; Cambridge, Massachusetts; London, England, 1998.

(62) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780.

(63) Coulom, R. Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search. In

Computers and Games, 5th International Conference; 2006; pp. 72–83.

(64) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(65) Weininger, D.; Arthur, W.; L., W. J. SMILES . 2 . Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.

(66) Yoshikawa, N.; Terayama, K.; Sumita, M.; Homma, T.; Oono, K.; Tsuda, K. Population-Based De Novo Molecule Generation, Using Grammatical Evolution. *Chem. Lett.* **2018**, *47*, 1431–1434.

(67) Maaten, L. Van Der; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.